

## PATENT ABSTRACTS OF JAPAN

(11) Publication number : 2001-236464

(43) Date of publication of application : 31.08.2001

(51) Int. Cl.

G06K 9/20  
G06T 7/00

(21) Application number : 2000-048603

(71) Applicant : RICOH CO LTD

(22) Date of filing : 25.02.2000

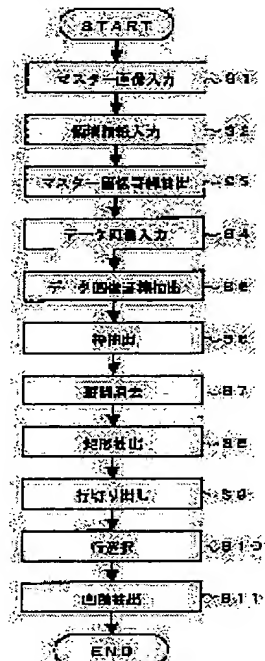
(72) Inventor : BESSHO GORO

(54) METHOD AND DEVICE FOR CHARACTER EXTRACTION AND STORAGE MEDIUM

(57) Abstract:

PROBLEM TO BE SOLVED: To accurately extract even a character which sticks out from a frame on a document.

SOLUTION: The frame is recognized (S6), and the ruled lines constituting the frame are erased (S7). Rectangles of black pixel connecting components in an area obtained by expanding the frame are extracted from the document image having the ruled lines erased (S8) and are integrated to extract a character string area (line) (S9). Only the character area, which can be considered to belong to the frame, is selected from the extracted character string area (S10), and its image is cut out of the document image having the ruled lines erased (S11).



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C): 1998, 2000 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-236464

(P2001-236464A)

(43) 公開日 平成13年8月31日 (2001.8.31)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テームト* (参考)
G 0 6 K 9/20	3 4 0	G 0 6 K 9/20	3 4 0 L 5 B 0 2 9
			3 4 0 K 5 L 0 9 6
G 0 6 T 7/00		G 0 6 F 15/70	3 3 0 Q

審査請求 未請求 請求項の数 7 O L (全 7 頁)

(21) 出願番号 特願2000-48603 (P2000-48603)

(22) 出願日 平成12年2月25日 (2000.2.25)

(71) 出願人 000006747

株式会社リコー

東京都大田区中馬込1丁目3番6号

(72) 発明者 別所 吾朗

東京都大田区中馬込1丁目3番6号 株式会社リコー内

(74) 代理人 100073760

弁理士 鈴木 誠 (外1名)

Fターム(参考) 5B029 AA01 BB02 CC28 EE12

5L096 BA17 CA24 EA15 EA35 FA03

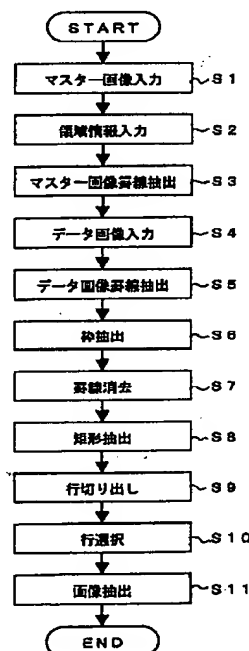
FA25 FA44 GA08 HA07

(54) 【発明の名称】 文字抽出方法、文字抽出装置及び記憶媒体

(57) 【要約】

【課題】 帳票上の枠からはみだした文字も正確に抽出する。

【解決手段】 枠を認識し (S 6)、その枠を構成する野線を消去する (S 7)。野線消去後の帳票画像より、枠を拡大した領域内の黒画素連結成分の矩形を抽出し (S 8)、それら矩形の統合によって文字列領域 (行) を抽出する (S 9)。抽出された文字列領域の中から枠に属するとみなし得る文字列領域だけを選択し (S 10)、その画像を野線消去後の帳票画像より切り出す (S 11)。



1

## 【特許請求の範囲】

【請求項 1】 2 値画像中の野線によって囲まれた枠を認識する第 1 ステップと、この第 1 ステップで認識された枠を構成する野線の黒画素を消去する第 2 ステップと、この第 2 ステップによって野線の黒画素を消去された 2 値画像から、前記第 1 ステップで認識された枠を拡大した領域の内部の黒画素連結成分の矩形を抽出する第 3 ステップと、この第 3 ステップによって抽出された矩形を統合して文字列領域を抽出する第 4 ステップと、前記第 1 ステップによって認識された枠に関連して前記第 4 ステップによって抽出された文字列領域の中で当該枠に属するとみなし得る文字列領域を選択する第 5 ステップとを含むことを特徴とする文字抽出方法。

【請求項 2】 前記第 5 ステップにおいて、文字列領域が枠に属するか否かを、当該文字列領域の当該枠の領域と重なる部分の面積割合に基づいて判定することを特徴とする請求項 1 記載の文字抽出方法。

【請求項 3】 前記第 3 ステップにおいて、枠の拡大量を当該枠の短辺の長さを基準にして決定することを特徴とする請求項 1 記載の文字抽出方法。

【請求項 4】 2 値イメージデータを入力する第 1 手段と、この第 1 手段によって入力された 2 値イメージデータから野線によって囲まれた枠を認識する第 2 手段と、この第 2 手段によって認識された枠を構成する野線の黒画素を前記入力された 2 値イメージデータより消去する第 3 手段と、この第 3 手段によって野線の黒画素が消去された後の 2 値イメージデータより、前記第 2 手段によって認識された枠を拡大した領域内の黒画素連結成分の矩形を抽出する第 4 手段と、この第 4 手段によって抽出された矩形を統合して文字列領域を抽出する第 5 手段と、前記第 2 手段によって認識された枠に関連して前記第 5 手段によって抽出された文字列領域の中で当該枠に属するとみなし得る文字列領域を選択する第 6 手段とを具備することを特徴とする文字抽出装置。

【請求項 5】 前記第 6 手段において、文字列領域が枠に属するか否かを、当該文字列領域の当該枠の領域と重なる部分の面積割合に基づいて判定することを特徴とする請求項 4 記載の文字抽出装置。

【請求項 6】 前記第 4 手段において、枠の拡大量を当該枠の短辺の長さを基準にして決定することを特徴とする請求項 4 記載の文字抽出装置。

【請求項 7】 請求項 1、2 又は 3 記載の文字抽出方法の各ステップの処理をコンピュータに実行させるためのプログラムが記録されたことを特徴とするコンピュータ読み取り可能な記憶媒体。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】 本発明は、画像処理の分野に係り、特に、文字認識装置などにおいて、表や帳票などの野線に囲まれた枠の内部に記入された文字を抽出する

2

技術に関する。

## 【0002】

【従来の技術】 帳票などの野線に囲まれた枠に記入された文字を抽出する方法として、特開平 9-161007 号公報に開示されているように、野線矩形を抽出し、野線矩形の外側の座標を用いて枠を認識し、野線画素を消去してから枠内の黒画素連結成分の矩形を抽出し、抽出した矩形の中で枠に接した矩形を除去し、残った矩形を用いて枠内の文字を切り出す方法が知られている。

## 【0003】

【発明が解決しようとする課題】 帳票において枠内に記入された文字が枠からはみ出していることがある。枠からはみ出した文字に対応した黒画素連結成分の矩形は、前記従来方法によれば枠に接触した矩形として除去されてしまう結果、枠からはみ出した文字を正常に抽出することができない。

【0004】 よって、本発明の目的は、枠からはみ出した文字も正常に抽出可能な文字抽出方法及び装置を提供することにある。

## 【0005】

【課題を解決するための手段】 上記目的を達成するため、本発明においては、2 値画像より野線によって囲まれた枠を認識し、認識した枠を構成する野線の黒画素を消去し、野線の黒画素を消去後の 2 値画像より、認識された枠を拡大した領域内の黒画素連結成分の矩形を抽出し、その矩形を統合して文字列領域を抽出し、認識された枠に関連して抽出された文字列領域の中で当該枠に属するとみなし得る文字列領域を選択する。また、文字列領域が枠に属するか否かの判定を、当該文字列領域の当該枠の領域と重なる部分の面積割合に基づいて行う。また、枠を拡大した領域の大きさを、枠の短辺の長さに応じて決定する。

【0006】 このような本発明の特徴及びその他の特徴について、実施の形態に関連して以下に詳述する。

## 【0007】

【発明の実施の形態】 以下、添付図面を参照し、本発明の実施の形態について説明する。図 1 は、本発明の実施の一形態である文字抽出装置のブロック構成の一例を示すブロック図である。図 2 は、この文字抽出装置における処理の流れを示すフローチャートである。また、また、図 4 乃至図 7 は処理の説明のため図である。

【0008】 この文字抽出装置は、例えば光学的文字認識装置の前処理部として用いられるもので、図 1 に見られるように、帳票の 2 値イメージデータを入力するためのスキャナなどの 2 値画像入力部 100 と、これによって入力された 2 値イメージデータを蓄積するためのマスターイメージメモリ 102 及びデータイメージメモリ 104 と、野線で囲まれた枠の認識に関連した領域情報入力部 106、領域情報ファイル 108、野線抽出部 110、アフィン変換係数メモリ 112、野線ファイル 11

3

4、野線メモリ116、枠抽出部118及び枠領域メモリ120と、認識された枠を構成する野線の黒画素を消去する（白画素に置き換える）ための野線消去部122と、枠の拡大領域内の黒画素連結成分の矩形を抽出するための矩形抽出部124と、抽出された矩形の情報を記憶するための矩形メモリ126と、矩形を統合して文字列領域（行）を抽出する行切り出し部128と、抽出された文字列領域の情報を記憶するための行メモリ130と、抽出された文字列領域の中で枠に属するとみなされる文字列領域を選択し、その2値イメージデータを切り出す画像抽出部132と、切り出された2値イメージデータを記憶するための文字領域画像メモリ134とから構成される。

【0009】以下、この文字抽出装置の動作について、処理の流れに沿って説明する。この文字抽出装置においては、帳票上の枠の認識のために、帳票のマスター画像（データの記入されていない空の帳票の画像）を利用するので、最初に2値画像入力部100によって帳票のマスター画像の2値イメージデータを入力し、それをマスターイメージメモリ102に格納する（図2のステップS1）。

【0010】そして、領域情報入力部106によって、マスター画像中の認識対象となる枠の位置や文字種などの情報を入力する（ステップS2）。具体的には、例えば、マスターイメージメモリ102内のマスター画像の2値イメージデータをディスプレイなどに表示し、マウスなどのポインティングデバイスを用いて認識対象の枠の位置などを指定する。入力された情報は領域情報ファイル108に格納される。なお、枠認識に利用されるのは、枠の位置に関する情報のみであるので、それ以外の情報の入力は省略可能である。

【0011】次に、野線抽出部110において、領域情報ファイル108内の枠の位置の情報を参照し、マスターイメージメモリ102内の2値イメージデータより認識対象の枠を構成する野線を抽出し、その情報（始点、終点の座標など）を野線ファイル114に格納する（ステップS3）。この野線抽出は、例えば、主走査方向及び副走査方向の所定値以上の長さの黒ランを抽出し、所定の距離の範囲内にある黒ランを矩形に統合することによって各方向の緯線矩形を抽出する一般的な方法によって行うことができる。

【0012】なお、ステップS1～S3は、新たなフォーマットの帳票を処理する場合にのみ行えばよいものである。例えば、同じフォーマットの帳票を多数枚連続して処理する時には、その1枚目の処理に先立ってステップS1～S3を行えばよい。この場合でも、そのフォーマットの帳票に関する領域情報及び野線情報が領域情報ファイル108及び野線ファイル114にそれぞれ登録されているならば、その情報を利用できるため、ステップS1～S3を改めて実行する必要はない。

4

【0013】以上のようなマスター画像に関する領域情報及び野線情報の登録が終了したならば（あるいは、既に得られている場合には、ステップS1～S3を行うことなく直ちに）、2値画像入力部100によって、帳票のデータ画像（データが記入された帳票の画像）の2値イメージデータが入力され、データイメージメモリ104に格納される（ステップS4）。

【0014】野線抽出部110において、この入力された帳票のデータ画像上に野線ファイル114に格納されている野線の情報を投影することにより、データ画像上の認識対象の枠を構成する野線を抽出し、その野線の情報（始点、終点の座標など）を野線メモリ116に格納する（ステップS5）。この際、マスター画像とデータ画像の間の位置ずれを計測し、その位置ずれの修正を行う。この位置ずれの計測には、例えば、両画像をシフトしながら、例えば野線が十字に交差する点などの特徴点について画像間でパターンマッチングを行い、最も良好に一致する相対位置を求めるような方法を利用できる。また、位置ずれの修正のためのアフィン変換係数が求められてアフィン変換係数メモリ112に記憶され、このアフィン変換係数が各野線の位置ずれの修正に適用される。データ画像の1枚1枚で位置ずれ量は異なるため、入力された各データ画像毎に、改めて位置ずれの計測とアフィン変換係数の算出が行われる。なお、位置ずれを問題にする必要がない場合には、位置ずれの計測と修正の処理を省略してよい。

【0015】次に、枠抽出部118において、野線メモリ116内の野線情報を参照することにより、野線によって囲まれた枠の領域を抽出し、その情報（始点と終点の座標など）を枠領域メモリ120に格納する（ステップS6）。

【0016】次に、野線消去部122において、野線メモリ116内の野線情報を参照し、抽出された野線の黒画素を白画素に置換する処理を、データイメージメモリ104内のデータ画像の2値イメージデータに対して施す（ステップS7）。

【0017】例えば、図4に示すような帳票のデータ画像上の「部番」の右側の枠が認識領域として指定された場合、図5に示すような枠150の領域が抽出され、また、この枠150を構成する野線は消去される（図4との対応関係を分かりやすくするため、図5においては野線の一部のみが消去されている）。

【0018】次に矩形抽出部124は、黒画素連結成分の矩形を抽出するが、認識された枠の内部について矩形抽出を行うのではなく、枠を拡大した領域の内部について矩形抽出を行い、その情報（始点、終点の座標など）を矩形メモリ126に格納する（ステップS8）。例えば、図5に示す枠150の場合、図6に示すように枠150を拡大した領域153の内部について矩形抽出を行う。枠の拡大量は枠の短辺のサイズ（枠150のheight）

5

t)を基準にして決定される。このようにすれば、枠もしくは文字の大きさの違いに適応できる。ここに示す例は、横書きの帳票であり縦方向に文字のはみ出しが生じやすいため、枠150の高さheightの半分だけ、枠150を上下に拡張している。

【0019】次に、行切り出し部128において、矩形メモリ126内の矩形情報を参照し、所定の距離範囲内にある矩形を統合することにより文字列領域(行)を抽出し、その情報(始点、終点の座標など)を行メモリ130に格納する(ステップS9)。日本語が用いられる帳票では、行の方向は縦方向と横方向の両方が考えられるが、多くの場合、帳票中の行はほとんどが横方向であるので、横方向に連続した矩形の統合により横方向の文字列領域を抽出すればよい。図6に示す枠の拡大領域153においては、図7に示す3つの文字列領域161、162、163が抽出される。

【0020】次に、画像抽出部132において、枠領域メモリ120内の枠領域情報及び行メモリ130内の文字列領域情報を参照し、各枠に関して抽出された文字列領域の中で、その枠に属するとみなし得る文字列領域を選択し(ステップS10)、選択した文字列領域の2値イメージデータをデータイメージメモリ104内の罫線消去後の2値イメージデータより切り出し、それを文字列領域画像メモリ134に格納する(ステップS11)。

【0021】文字列領域が枠に属するとみなし得るか否かの判定は、例えば次のようにして行われる。

(1) 拡大前の枠の内部に全体が包含される文字列領域は、枠に属する文字列領域と判定する。

(2) 全体が拡大前の枠の外にある文字列領域は、枠に属する文字列領域ではないと判定する。

(3) 上記条件のいずれにも該当しない文字列領域については、文字列領域の拡大前の枠の領域と重なる部分の、文字列領域全体に対する面積割合を計算する。この面積割合が1/2以上(全体の半分以上が拡大前の枠と重なる)文字列領域は枠に属する文字列領域と判定し、面積割合が1/2未満の文字列領域は枠に属する文字列領域ではないと判定する。

【0022】図7の文字列領域161、162、163のうちで、文字列領域161は上記条件(1)により枠150に属すると判定される。文字列領域163は、上記条件(2)により枠150に属しないと判定される。文字列領域162は、面積割合が1/2未満であるため、上記条件(3)により枠150に属しないと判定される。したがって、枠150の文字列領域としては文字列領域161だけが選択され、その画像が正しく切り出される。また、このような文字列(行)を単位として、枠に属するか否かの判定を行うことにより、分離文字の一部が誤って抽出されたり、抽出されなかったりする不都合を回避できる。枠150に属しないと判定される文字列領域162は、1つ下の「品名」の枠の処理時に、

6

その枠に属する文字列領域として選択され、その2値イメージデータが切り出されることは以上の説明から明らかである。

【0023】このように、文字列領域162内の文字のように、枠からはみ出した文字も正しく抽出することができる。なお、枠に属する文字列領域が2つ以上抽出される場合には、それら文字列領域の画像を1つの文字列領域の画像として統合して抽出するようにしてもよい。

【0024】本発明の他の実施の形態によれば、マスター画像に関して予め登録した情報を利用せず、帳票のデータ画像から直接的に枠が認識される。図1を援用して説明すれば、罫線抽出部110において、例えば、データイメージメモリ104内の2値イメージデータをスキャンし、主走査方向及び副走査方向の所定値以上の長さの黒ランを抽出し、抽出した各方向の黒ランに対し、所定の距離範囲内にある黒ランを統合する処理を行うことによって罫線の矩形を抽出し、その始点と終点の座標を罫線メモリ116に格納する。以下、前述の実施の形態の場合と同様の処理を行って枠の文字の抽出を行う。なお、領域情報入力部106と同様に、帳票のデータ画像をディスプレイに表示し、マウスなどを用いて認識対象の枠の位置を指定し、指定された枠を構成する罫線のみを抽出するようにしてもよい。このような実施の形態は、予め形式の定まっていない不定形の帳票を処理する場合に有効である。

【0025】前述のような本発明による文字抽出装置及びその処理は、必ずしも専用のハードウェアによって実現される必要はなく、例えば図3に示すようなCPU201、メモリ202、ハードディスク203、入力装置(イメージスキャナ、キーボード、マウスなど)204、ディスプレイ206、各種記憶媒体(磁気ディスク、光ディスク、光磁気ディスク、メモリカードなど)207の読み書きのための媒体ドライブ205、外部の機器やネットワークとの通信のための通信装置(モデム、ネットワークアダプタなど)208などをバス210で接続した一般的な構成のコンピュータを利用して、ソフトウェアにより実現することもできる。この場合、文字抽出装置の各部の機能をコンピュータ上で実現するためのプログラム、換言すれば、文字抽出のための各処理ステップをコンピュータ上で実行するためのプログラムが、例えば、それが記録された記憶媒体207から媒体ドライブ205によってメモリ202に読み込まれたり、外部装置より通信装置208によってメモリ202に読み込まれたり、あるいはハードディスク203からメモリ202に読み込まれ、CPU201により実行される。このプログラムを、それを固定記憶させた半導体ROMとして実装してもよい。このようなプログラムが記録された各種記憶媒体207や半導体ROMなども本発明に包含されるものである。また、帳票の2値イメージデータは、例えば入力装置204に含まれるスキャナ

7

によって読み込まれたり、記憶媒体 207 から読み込まれたり、あるいは通信装置 208 によって外部機器から入力される。

#### 【0026】

【発明の効果】以上の詳細な説明から明らかなように、本発明の文字抽出方法又は装置によれば、枠から文字がはみ出して記入されている場合であっても、正確な文字抽出が可能となる。したがって、本発明の文字抽出方法又は装置を文字認識の前処理に適用すれば、枠からはみ出して記入された文字も正確な文字認識が可能となる。また、本発明の記憶媒体を用いれば、一般的なコンピュータを利用して上に述べたような正確な文字抽出が可能となる、等々の効果を得られる。

#### 【図面の簡単な説明】

【図1】本発明による文字抽出装置のブロック構成の一例を示すブロック図である。

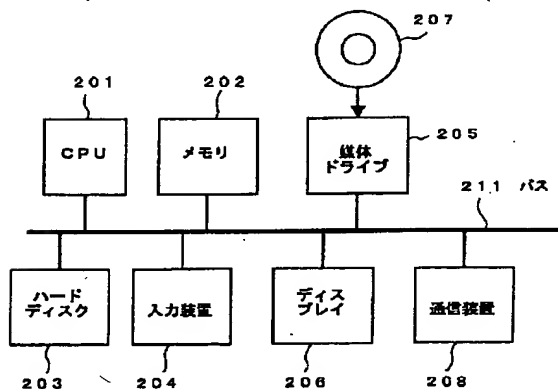
【図2】図1の文字抽出装置における処理の流れを示すフローチャートである。

【図3】本発明をソフトウェアにより実施するために利用されるコンピュータの一例を示すブロック図である。

【図4】帳票のデータ画像と、その枠の一例を示す図である。

【図5】罫線消去後のデータ画像上の枠の領域を示す図である。

【図3】



8

【図7】枠を拡大した領域から抽出される文字列領域を示す図である。

#### 【符号の説明】

- 100 2値画像入力部
- 102 マスターイメージメモリ
- 104 データイメージメモリ
- 106 認識領域情報入力部
- 108 領域情報ファイル
- 110 罫線抽出部
- 112 アフィン変換係数メモリ
- 114 罫線ファイル
- 116 罫線メモリ
- 118 枠抽出部
- 120 枠領域メモリ
- 122 罫線消去部
- 124 矩形抽出部
- 126 矩形メモリ
- 128 行切り出し部
- 130 行メモリ
- 132 画像抽出部
- 134 文字領域メモリ
- 150 枠
- 153 枠の拡大領域
- 161, 162, 163 文字列領域

【図4】

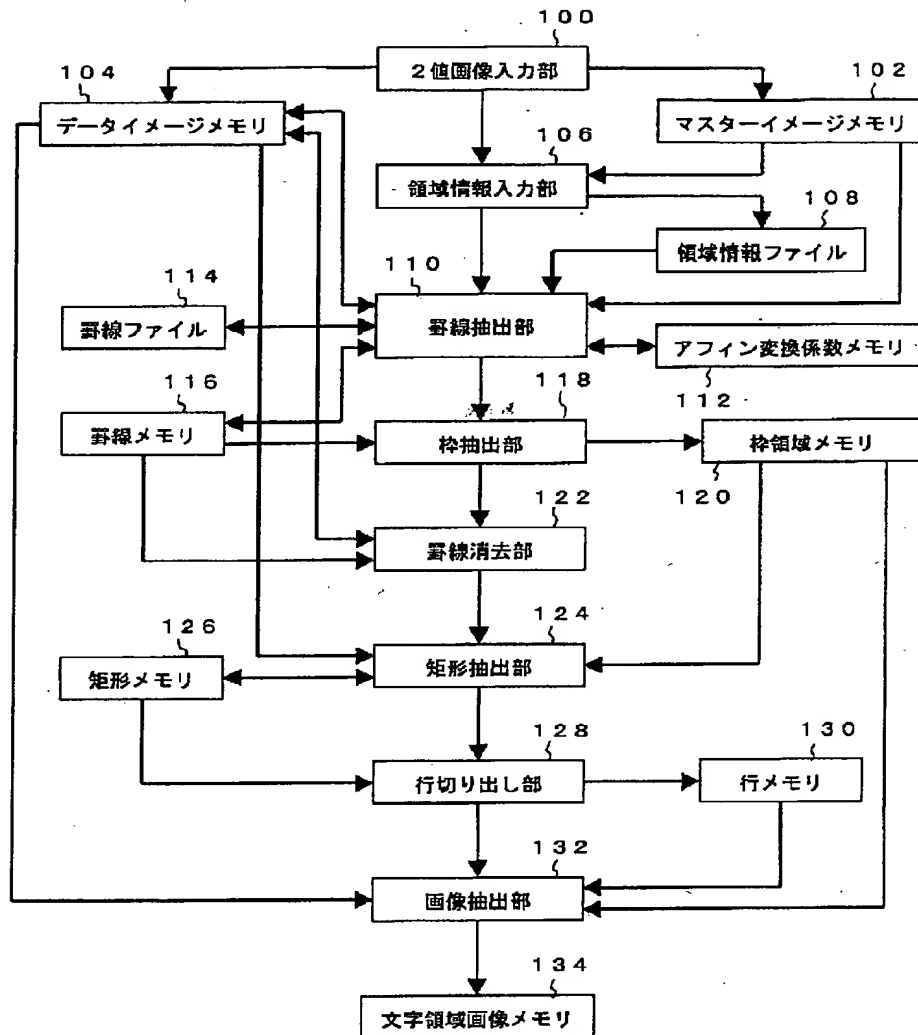
番	規	年	月	日
部 番	A2291078			
品 名	コティン <sup>ン</sup> サタ <sup>ン</sup> リ			
代表機種	A228			

【図5】

番	規	年	月	日
部 番	A2291078			
品 名	コティン <sup>ン</sup> サタ <sup>ン</sup> リ			
代表機種	A228			

150 枠領域

【図1】



【図6】

番	規	年	月	日
部 番	A2291078			
品 名	コティンソソクン			
代表機種	A229			

153 拡大領域

【図7】

番	規	年	月	日
部 番	A2291078			
品 名	コティンソソクン			
代表機種	A229			

153

【図 2】

